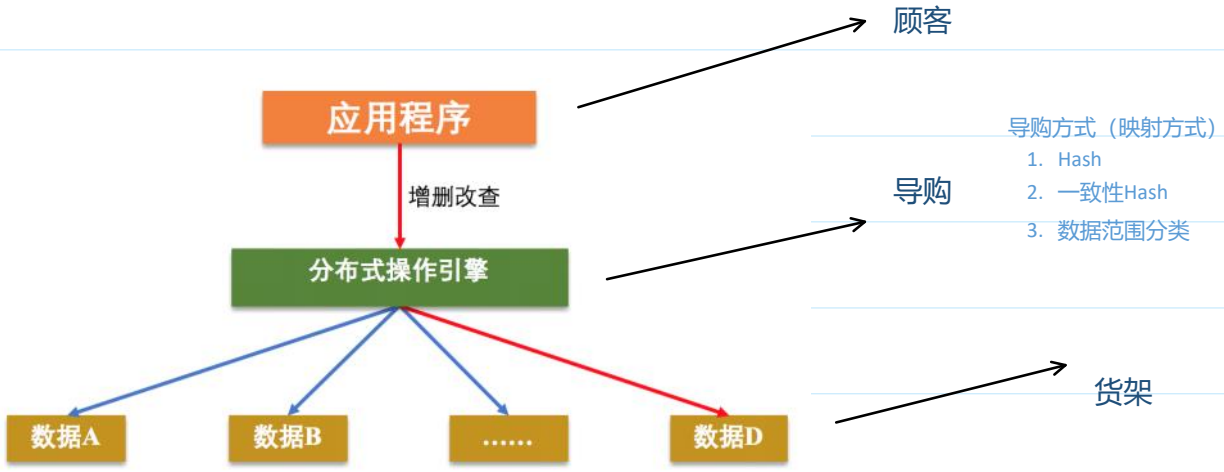
5.2 分布式数据存储系统三要素

2022年3月14日 15:44



顾客：生产和消费数据

数据分类

1. 结构化数据
2. 半结构化数据
3. 非结构化数据

- 结构化数据通常是指关系模型数据，其特征是数据关联较大、格式固定。火车票信息比如起点站、终点站、车次、票价等，就是一种结构化数据。结构化数据具有格式固定的特征，因此一般采用分布式关系数据库进行存储和查询。
- 半结构化数据通常是指非关系模型的，有基本固定结构模式的数据，其特征是数据之间关系比较简单。比如 HTML 文档，使用标签书写内容。半结构化数据大多可以采用键值对形式来表示，比如 HTML 文档可以将标签设置为 key，标签对应的内容可以设置为 value，因此一般采用分布式键值系统进行存储和使用。
- 非结构化数据是指没有固定模式的数据，其特征是数据之间关联不大。比如文本数据就是一种非结构化数据。这种数据可以存储到文档中，通过 ElasticSearch（一个分布式全文搜索引擎）等进行检索。

导购：确定数据位置

数据分片技术 → 分布式存储系统按照一定的规则将数据存储到相对应的存储节点中

货架：存储数据

货架分类

1. 分布式数据库
2. 分布式键值系统
3. 分布式存储系统

- 分布式数据库，通过表格来存储结构化数据，方便查找。常用的分布式数据库有 MySQL Sharding、Microsoft SQL Azure、Google Spanner、Alibaba OceanBase 等。
- 分布式键值系统，通过键值对来存储半结构化数据。常用的分布式键值系统有 Redis、Memcache 等，可用作缓存系统。具体的缓存技术我将在第 27 篇文章“分布式数据之缓存技术：‘身手钥钱’随身带”中与你详细介绍。
- 分布式存储系统，通过文件、块、对象等来存储非结构化数据。常见的分布式存储系统有 Ceph、GFS、HDFS、Swift 等。

分布式数据库	MySQL Sharding	SQL Azure	Spanner	OceanBase
起源	瑞典MySQL AB公司开发，目前属于Oracle公司	微软在SQL Server技术基础上发展出来的云端关系型数据库服务	Google研发的可扩展的、全球分布式的数据库	阿里研发的高性能分布式数据库系统
是否开源	是	否	否	是
数据格式	结构化数据	结构化数据	结构化数据	结构化数据
应用场景	最流行的关系型数据库管理系统；在Web应用方面，是最好的关系数据库管理系统应用软件之一	云端数据库平台，适用于云端存储大规模结构化数据的场景	全球分布式的数据库，适用于全球性、大规模的结构化数据存储场景	支持海量数据存储和操作，比如 实现了上千亿条记录和上百TB数据上的跨行跨表事务，目前用于 用于存储淘宝具体的商品、店铺等信息

分布文件存储系统	Ceph	GFS	HDFS	Swift
起源	最初起源于塞奇 · 韦伊（Sage Weil）就读博士期间的工作	Google的分布式文件存储系统	Hadoop的核心子项目，为类似Hadoop这样的云计算而生	最初由Rackspace公司开发，2010 年贡献给 OpenStack 开源社区
是否开源	是	否	是	是
系统架构	去中心化	集中式架构	集中式架构	去中心化
数据格式	非结构化数据	非结构化数据	非结构化数据	非结构化数据
应用场景	通用的实时存储系统，适合频繁读写场景	基于Linux的专有大规模分布式文件系统，大文件读写场景	大数据场景，擅长处理离线批量大数据；	openstack对象存储场景


分布式存储系统三要素
顾客
定义
顾客可以生产和消费数据，相当于分布式存储系统中的应用程序
顾客分类
生产者
给存储系统添加数据
消费者
使用系统中存储的数据
数据分类
结构化数据
数据关联较大、格式固定
半结构化数据
数据之间关系比较简单，有基本固定的结构
非结构化数据
数据之间关联不大，没有固定模式
导购
数据分片
定义
相当于一种导购技术，分布式存储系统按照一定的规则将数据存储到相对应的存储节点中，或者到相对应的存储节点中获取想要的数据
作用
降低单个存储节点的存储和访问压力
快速找到数据所在的存储节点，大大降低搜索延迟，提高用户体验
方法
数据特征分片
数据范围分片
哈希分片
一致性哈希分片
……
数据复制
定义
将数据进行备份，多个节点存储相同的数据
作用
主备方式存储，保证数据的可靠性
货架
定义
货架是用来存储数据的
分类
分布式数据库
作用
通过表格来存储结构化数据
实例
MySQL Sharding、Microsoft SQL Azure、Google Spanner、Alibaba OceanBase ……
分布式键值系统
作用
通过键值对来存储半结构化数据，可用作缓存系统
实例
Redis、Memcache ……
分布式文件系统
作用
通过文件来存储非结构化数据
实例
Ceph、GFS、HDFS、Swift ……